PROCEEDINGS B

Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

Journal:	*Proceedings B*
Manuscript ID	RSPB-2016-2290.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Puttick, Mark; University of Bristol, Earth Sciences O'Reilly, Joseph; University of Bristol, School of Earth Sciences Tanner, Alastair; University of Bristol, Biological Sciences Fleming, James; University of Bristol, School of Earth Sciences Clark, James; University of Bristol, School of Earth Sciences Holloway, Lucy; University of Bristol, School of Earth Sciences Lozano-Fernandez, Jesus; University of Bristol, School of Earth Sciences Parry, Luke; University of Bristol, Life Sciences Building Tarver, James; University of Bristol, Earth Sciences Pisani, Davide; University of Bristol, School of Earth Sciences Donoghue, Philip; University of Bristol, School of Earth Sciences
Subject:	Evolution < BIOLOGY, Palaeontology < BIOLOGY, Taxonomy and Systematics < BIOLOGY
Keywords:	phylogeny, Bayesian, parsimony, morphology, palaeontology, Maximum Likelihood
Proceedings B category:	Palaeobiology

Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

Mark N. Puttick[1,3]*, J. E. O'Reilly[1]*, Alastair R. Tanner[2], James F. Fleming[1], James Clark[1], Lucy Holloway[1], Jesus Lozano-Fernandez[1,2,] Luke A. Parry[1], James E. Tarver[1], Davide Pisani[1,2], Philip C.J. Donoghue[1†]

[1]School of Earth Sciences, [2]School of Biological Sciences, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK, [3]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

*These authors contributed equally to this study

†Corresponding authors: davide.pisani@bristol.ac.uk; phil.donoghue@bristol.ac.uk

Abstract

Morphological data provides the only means of classifying the majority of life's history, but the choice between competing phylogenetic methods for the analysis of morphology is unclear. Traditionally, parsimony methods have been favoured but recent studies have shown that these approaches are less accurate than the Bayesian implementation of the Mk model. Here we expand on these findings in several ways: we assess the impact of tree shape and maximum-likelihood estimation using the Mk model, as well as analysing data composed of both binary and multistate characters. We find that all methods struggle to correctly resolve deep clades within asymmetric trees, and when analysing small character matrices. The Bayesian Mk model is the most accurate method for estimating topology, but with lower resolution than other methods. Equal weights parsimony is more accurate than implied weights parsimony, and Maximum Likelihood estimation using the Mk model is the least accurate method. We conclude that the Bayesian implementation of the Mk model should be the default method for phylogenetic estimation from phenotype datasets, and we explore the implications of our simulations in reanalysing several empirical morphological character matrices. A consequence of our finding is that high levels of resolution or the ability to classify species or groups with much confidence should not be expected when using small datasets. It is now necessary to depart from the

traditional parsimony paradigms of constructing character matrices, towards datasets constructed explicitly for Bayesian methods.

Keywords: phylogeny, Bayesian, parsimony, cladistics, morphology, palaeontology

1. Introduction

The fossil record affords the only direct insight into evolutionary history of life on Earth, but the incomplete preservation and temporal distribution of fossils has long prompted biologists to seek alternative perspectives, such as molecular phylogenies of living species, eschewing palaeontological evidence altogether [1]. However, there is increasing acceptance that analyses of historical diversity cannot be made without phylogenies that incorporate fossil species [2,3] and calibrating molecular phylogenies to time cannot be achieved effectively without recourse to the fossil record [4]. Integrating fossil and living species has become the grand challenge and there has been a modest proliferation of phylogenetic approaches to the analysis of phenotypic data. While conventional parsimony remains the most widely employed method, alternative parsimony [5] and probabilistic [6] models have been developed to better accommodate heterogeneity in the rate of evolution among characters and across phylogeny. Unfortunately, these competing methods invariably yield disparate phylogenetic hypotheses among which it is difficult to discriminate since the true tree is never known for empirical data.

A number of studies have attempted to establish the efficacy of competing phylogenetic methods using data simulated from known trees [7–9], finding that the probabilistic Mkv model outperforms parsimony methods, among which, conventional equal-weights parsimony performs best. However, these studies were potentially biased by their experimental design: (i) two of the studies employed a generating tree that was unresolved and, therefore, biased against parsimony methods which recover resolved trees; (ii) these studies did not discriminate between the impact of the probabilistic model and its implementation in a Bayesian framework; (iii) based on single empirical trees, the impact of tree symmetry, which is known to confound phylogeny estimation [10], was not explored; (iv) only binary characters were considered, whereas empirical datasets are commonly a mixture of binary and multistate characters. Therefore, we compare the performance of equal-weights parsimony (EW-Parsimony), implied-weights parsimony (IW-Parsimony), Maximum Likelihood and Bayesian implementations of the Mk model, based on datasets with different numbers of characters, comprised of binary and multistate characters and simulated on a fully balanced and a maximally-imbalanced phylogenetic tree. We find

that Bayesian inference out-performs all other methods, while EW-Parsimony performs better than IW-Parsimony, and Maximum Likelihood performs worst of all. We apply these competing phylogenetic methods to empirical morphological datasets of similar sizes to our simulated datasets and explore the efficacy of the ensuing phylogenetic hypotheses in light of the conclusions derived from of our simulation-based study.

2. Materials and Methods

(a) Simulation of Morphological Matrices

We simulated data on two 32-taxon generating trees at the extremes of tree symmetry: one fully asymmetrical and one fully symmetrical (see Supplementary Fig S1). For each tree we simulated matrices of three sizes: 100, 350 and 1000 characters. We generated matrices using the HKY + Γ Continuous model of molecular substitution, with $\kappa = 2$, the shape (set equal to rate) of the gamma distribution and underlying substitution rate for each replicate sampled from independent and identically distributed exponential distributions with a mean of one, and character state stationary frequencies fixed as $\pi = [0.2,0.2,0.3,0.3]$. We used a fixed and uneven stationary distribution of nucleotide frequencies to ensure our simulation model did not collapse into the Mk model, as this would bias the analysis in favour of Mk model-based approaches. We simulated 1000 replicate matrices with unique substitution parameters for each tree and each character number, resulting in a total of 6000 matrices. We set two types of character within each matrix, binary and multistate, and we simulated a proportion of 55 binary:45 multistate characters, based on the mean ratio found in a survey of empirical morphological data matrices [11]. We established binary characters by converting data simulated under the HKY model to R/Y coding (i.e. 0/1): morphological multistate characters were simulated by converting DNA bases to integers.

To ensure that our simulated data are realistic, we generated each set of 1000 unique replicate matrices such that the among-matrix distribution of homoplasy approximated the distribution of empirical homoplasy, characterised by the Consistency Index (CI), reported by Sanderson and Donoghue [12]. To approximate this distribution of homoplasy we placed the Sanderson and Donoghue data into

quantised bins of CI spanning 0.05, between the empirical bounds of 0.26 and 1.0, and simulated matrices until we matched this expected density per bin (Supplementary Figure S2).

The code used to simulate these data is available in Electronic Supplementary Materials.

(b) Phylogenetic analysis

We analysed the simulated matrices with EW-Parsimony, IW-Parsimony (k=2), and the Mk model [6] under both Maximum Likelihood and Bayesian implementations. EW-Parsimony and IW-Parsimony estimation of topology was performed in TNT [13]. We used the Mk plus gamma model for Maximum Likelihood estimation of topology in RAxML 7.2 [14], and Bayesian estimation of topology in MrBayes 3.2 [15]. As the approximate likelihood calculation of RAxML may be distant from the true likelihood [16], we conducted a sensitivity test by re-analysing a subset of our data with the likelihood implementation of the MK model in IQ-tree [17]; both methods gave effectively identical results, indicating results from the likelihood MKv model are not software-specific.

The Mkv model is inappropriate due to the lack of acquisition bias in the simulated data. For both Maximum Likelihood and Bayesian analyses we applied the discretised gamma distribution model to account for between-character rate heterogeneity. For Bayesian analyses the posterior distribution was sampled one million times by four chains using the Metropolis-coupled Markov-chain Monte Carlo algorithm, two independent runs were performed for each replicate and the two resulting posterior samples were combined after qualitative assessment of convergence. For parity, we characterised the result of all phylogenetic methods as the majority-rule consensus of resultant tree samples. We did not employ bootstrap methods to measure support for parsimony and likelihood analyses because phenotypic data does not meet the assumption that phylogenetic signal is distributed randomly among characters.

We used the Robinson-Foulds metric [18] to compare the similarity of estimated topologies against their respective generating tree. We also noted the per-node resolution, and the variation of node accuracy across the topology.

(c) Empirical analyses

We analysed four published palaeontological phenotype character matrices that encompass a range of character numbers and a diverse sample of taxa from the Tree of Life [19–22]. We resolved any ambiguities in character coding to their most derived state for each matrix to make analyses compatible across the different phylogenetic methods, facilitating comparison of results. We analysed each matrix by applying the same settings used to analyse our simulated matrices: EW-Parsimony, IW-Parsimony, as well as Bayesian and Maximum Likelihood implementations of the Mk model. Empirical morphological matrices are rarely constructed to contain invariant or parsimony uninformative characters. Therefore, the Mkv extension of the Mk model, which uses conditional likelihood to correct for such acquisition biases, is more appropriate than the Mk model for analysis of these empirical data matrices [6].

3. Results

(a) Simulated data

Accuracy is higher for trees inferred from data simulated on a symmetrical topology compared to trees estimated from data simulated on the asymmetrical topology (compare figures 2, 3). Bayesian consensus phylogenies are generally the least well resolved (Figure 1). All methods estimated topologies with greater accuracy as the number of analysed characters increased (figs 2, 3; supplementary table S5-7). All methods, apart from Maximum Likelihood, produced phylogenies with greater resolution with higher numbers of characters (figure 1).

For all implementations and dataset sizes, the Bayesian implementation of the Mk model achieves higher accuracy compared to other methods (table 1; figs 1-3). The two parsimony methods achieved the next highest levels of accuracy, EW-Parsimony achieving greater accuracy than IW-Parsimony. Maximum Likelihood was the least accurate method for topology reconstruction for both the symmetrical and asymmetrical phylogenies (table 1). The relative accuracy of these phylogenetic methods remains the same across all dataset sizes and the two simulation topologies (table 1; figs 1-3).

Nodes closer to the tips are significantly more accurately reconstructed in the asymmetrical phylogenies across all dataset sizes (table 2; figure 2; supplementary figure S8). In the symmetrical trees, there was no significant correlation between distance from the tips and the accuracy of node reconstruction, except in the Maximum Likelihood analysis of 100 characters (figure 2; table 2).

(b) Empirical phylogenies

Patterns of resolution achieved from the simulated datasets are similar for the empirical datasets. The Bayesian implementation of the Mk model estimates the least resolved phylogenies and Maximum Likelihood produces fully-resolved trees (full trees are shown supplementary figure S9-15).

Kulindroplax, from the Sutton et al. [22] dataset, is supported as a crown-mollusc based on Maximum Likelihood, EW-Parsimony, and IW-Parsimony (figure 4a-d). The results of the IW-Parsimony analysis are most similar to the original results [22], with *Kulindroplax* resolved as a crown-aplacophoran; Maximum Likelihood analysis of the dataset resolved *Kulindroplax* as the stem-aplacophoran. The result of the Bayesian analysis of the dataset is largely unresolved, and *Kulindroplax* is not discriminated as a member of any clade within molluscs or even as a member of total-group Mollusca.

The anthophyte hypothesis (non-monophyletic gymnosperms sister to seed ferns plus angiosperms) recovered by Hilton and Bateman [19] is supported by our EW-Parsimony and Maximum Likelihood analyses of their dataset which recovered a paraphyletic seed ferns plus Gnetophyta as sister to angiosperms (figure 4f, g); the results of Bayesian and IW-Parsimony analyses of the same dataset contradict the anthophyte hypothesis (figure 4e, h). The Bayesian analysis produced a non-monophyletic gymnosperms with the relationships between them and seed ferns unresolved with the exception of *Bennettitales* which resolved as a gnetophyte, and *Caytonia* as sister to the angiosperms.

Analyses of the Luo et al. [20] dataset yielded congruent results with the original study, with the placement of *Haramiyavia* outside of crown Mammalia and

multituberculates, although some haramiyids are resolved as crown mammals in the IW-Parsimony analysis (figure 5a-d).

Nyasasaurus is recovered as a member of Dinosauria in the Maximum Likelihood, EW-Parsimony, and IW-Parsimony analyses of the dataset from Nesbitt et al. [21] (figure 5e-h). The Bayesian analysis recovers *Nyasasaurus* in a polytomy with the two major clades of dinosaurs, corroborating the conclusion of Nesbitt et al. [21] that, given the data, its precise phylogenetic position is uncertain.

4. Discussion

(a) Simulations indicate that the Bayesian implementation of the Mk model outperforms all other methods and implementations

Previous simulation-based analyses that have attempted to evaluate the performance of likelihood and parsimony-based phylogenetic methods for analysing phenotypic data have found that the probabilistic model performs best [7,8]. However, these studies were biased against parsimony because they employed an unresolved generating tree that is problematic since parsimony methods will attempt to recover a fully resolved tree from the simulated data yielding a non-zero RF distance from the generating tree, even if the two trees are effectively compatible. Furthermore, since previous simulation studies considered the Mk model only within a Bayesian framework, they did not distinguish between the impact of the probabilistic model of character evolution and the statistical framework in which it was implemented.

Our analyses control for these shortcomings of previous simulation studies and show consistently that the Bayesian implementation of the Mk model performs best. In line with previous simulations [8], we found that EW-Parsimony performs better than IW-Parsimony. There is overlap between model performance shown by the distribution of Robinson-Foulds distances (table 1), but there is reason to have different degrees of confidence in the models; only the Bayesian implementation produces a relatively small distribution of tree performance compared to the large tails signifying worse performance in the two parsimony methods (table 1). We also found that the Bayesian implementation of the Mk model outperforms the Maximum Likelihood implementation, indicating that it is not merely the probabilistic transition model that

outperforms parsimony methods, but the implementation of the Mk model within a Bayesian statistical framework. Indeed, the Maximum Likelihood implementation of the Mk model was the worst-performing method, worse even than IW-Parsimony. In part, the poor performance of the Maximum Likelihood-Mk method is because we did not capture phylogenetic uncertainty associated with this phylogenetic method. This is normally achieved in analyses of molecular datasets through bootstrapping methods, but these are inappropriate for the analysis of phenotypic data since the basic methodological assumption, that the phylogenetic signal is randomly distributed across sites (characters), is not true for morphological data.

However, irrespective of the phylogenetic method used, dataset size correlated positively with both phylogenetic accuracy and resolution, diminishing differences in the relative performance of the competing phylogenetic methods. All phylogenetic methods also performed best when attempting to recover a symmetrical target tree; all methods found recovery of asymmetrical trees challenging and phylogenetic accuracy diminished from tip to root. The impact of tree topology is of particular concern since empirical phylogenetic trees are invariably asymmetric [23], and trees of fossil species are infamous for their asymmetry [24,25]. However, there is a broad spectrum of tree symmetry, with fully symmetric and fully asymmetric trees representing end-members. Palaeontological trees with the dimensions used in our simulations are typically far from the fully asymmetric pectinate generating tree we employed (Ic = ~0.4 for 32 species) [25]. Furthermore, the asymmetry of many palaeontological trees is often a representational artefact of attempting to summarise character evolution, or an analytic artefact of analysing the relationships among diverse clades based on representative species or higher taxa [26]. Thus, the challenge of recovering trees of extinct taxa may not be as great as a simplistic interpretation of our results might suggest.

(b) Analyses of empirical data bear out conclusions based on simulations

Maximum Likelihood, IW-Parsimony and EW-Parsimony methods of the simulated datasets commonly identify a single optimal tree, but the differences between the optimal trees derived from these methods provides no confidence that any one of the inferred topologies is accurate with reference to the placement of a taxon of interest. This view is corroborated by our reanalysis of empirical datasets which recovered

poorly resolved trees using the Bayesian implementation of the Mk model, and in a number of instances, indicate that the conclusions drawn in the corresponding original studies are not supported by the data.

In an extreme example, our re-analyses of the dataset published by Sutton et al. [22], which attempted to demonstrate a crown-aplacophoran mollusc affinity for *Kulindroplax*, yielded disparate hypotheses of affinity. EW-Parsimony and IW-Parsimony recovered the published result, while Maximum Likelihood recovered *Kulindroplax* as a stem-aplacophoran, and Bayesian could not discriminate *Kulindroplax* as a total-group mollusc (figure 4a). This poor resolution is unlikely to be a result of poor fossil evidence but, rather, the lack of discriminatory power in the small character matrix. Among the analyses of the dataset from Hilton and Bateman [19] we recovered some of the principal competing topologies that have featured in debate over the affinity of seed plants in past decades. However, the Bayesian analysis of the dataset recovered a topology that is largely unresolved in terms of the relationships among key clades. This suggests that the available data are insufficient to discriminate among the competing hypotheses, and this long-standing debate is largely an artefact of the false resolution of parsimony methods.

Bayesian analyses need not overturn the results from previous analyses based on deterministic phylogenetic methods like EW-Parsimony, IW-Parsimony, and Maximum Likelihood. A phylogenetic position for haramiyids, outside crown-Mammalia, is corroborated by our Bayesian analysis of the dataset from Luo et al. [20] - in contrast to the crown-Mammalia affinity recovered for some haramiyids through IW-Parsimony analysis of the same data (figure 5d). Similarly, *Nyasasaurus* was posited as the earliest dinosaur, and this conclusion is supported by the Bayesian analyses (Figure 5e) although this is not supported by EW-Parsimony, IW-Parsimony, and Maximum Likelihood analyses (Figure 5f-h). However, the Bayesian analysis is more robust in expressing the phylogenetic ambiguity identified by the original authors [19], as *Nyasasaurus* falls in a polytomy alongside the two major clades of dinosaurs.

Some of the differences between methods may simply reflect the dimensions of the dataset. The two datasets that cannot resolve relationships under Bayesian inference

and exhibit significant topological discordance among phylogenetic methods [19,22], are both comparatively small (34 taxa, 48 characters and 48 taxa, 82 characters, respectively). These both fall within the scope of simulated datasets that yield low resolution from the Bayesian method and, from other phylogenetic methods, high resolution but low accuracy (figure 1). The two empirical datasets that yield trees with greater congruence from the different phylogenetic methods, are both larger: Luo (114 taxa, 497 characters) and Nesbitt (82 taxa, 413 characters). The size of these matrices is comparable with our simulation results in which we see marked increases in topological accuracy and agreement between methods (figure 1 - between 350 and 1000 characters).

(c) Implications for phylogenetic analysis of phenotypic data

The results of our simulation studies indicate that the cadre of phylogenetic hypotheses generated from phenotypic data using parsimony methods require reassessment using the Bayesian implementation of the Mk model. It is likely that many evolutionary interpretations are contingent on precise but inaccurate phylogenetic hypotheses. In this undertaking it is important that the implications of our simulation studies are considered in the design of phylogenetic studies.

Firstly, phylogenies of fossils tend toward strong asymmetries [25] and, like all phylogenetic methods, Bayesian inference struggles with the recovery of deep nodes within asymmetric trees. Therefore, it is important that outgroups are sampled extensively, ensuring that contentious in-group relationships are closer to the tips, where topological accuracy is highest. Further, in-group lineages should be sampled in a manner that does not accentuate tree asymmetry.

Secondly, phylogenetic accuracy and resolution correlates positively with the relative dimensions of the dataset. Accordingly, phylogenetic resolution or certainty should not be expected from cladistic analyses of small morphological datasets (i.e, those around 100 characters or fewer), particularly if they include fossils. There are finite limits to the number of available phylogenetically-informative characters [27] and, for well-studied clades, it may be perceived that these phylogenetically informative characters have already been found. However, it is important to note that the concept of phylogenetic informativeness is different within a likelihood versus a parsimony

framework: in parsimony characters that undergo few changes are prized in favour of homoplastic characters. Under the likelihood model, branch length, informed by the number of character changes, contributes to topology estimation. Thus, traditionally 'bad' phylogenetic characters (those exhibiting homoplasy) may find utility in expanding the dimensions of phenotypic character matrices as long as homoplasy falls within the limits that the model can accommodate. In a Bayesian framework this can be tested using posterior predictive tests of model adequacy (e.g. [28]).

Finally, we may need to alter our expectations to anticipate less well-resolved but more accurate phylogenetic hypotheses, which will both constrain and guide research. Greater resolution may be found by generating matrices suited to likelihood- rather than parsimony-based phylogenetic methods. However, we must also come to terms with the prospect that for some groups of organisms, or their fossil remains, there may be insufficient data. As such, their evolutionary relationships might not therefore be resolvable using morphological data alone and, if they are fossils, their evolutionary significance may never be realised. Nevertheless, resolving phylogenies is not the end game for evolutionary biology. Incompletely resolved trees can still be used as a basis for investigating interesting macroevolutionary questions, and methods exist for incorporating tree uncertainty in phylogenetic comparative methods [e.g [29]).

5. Conclusions

A growing consensus shows that the Bayesian Mk model is the most accurate method of phylogenetic reconstruction, and here we show that this remains true across dramatically different tree shapes, when analysing datasets composed of both multistate and binary characters, and when compared to Maximum Likelihood estimation using the Mk model. We would recommend that Bayesian implementations of the Mk model should become the default method for phylogenetic analyses of cladistic morphological datasets, and we should expect low levels of resolution with small datasets. As parsimony methods appear to be less effective than probabilistic approaches, it may be necessary to alter data collection practices by moving away from choosing a selection of characters that undergo few changes, and moving towards scoring all possible characters from the available taxa irrespective of their expected homoplasy.

Data Accessibility. The code used to simulate the data used in this publication can be accessed from Dryad at doi:10.5061/

Author's Contributions. All authors contributed to the design of the study; MNP and JoR led the analyses; interpretation of results and writing was led by MNP, JoR, PD, and DP, though all authors contributed to the interpretation of results and the writing of the manuscript.

Competing Interests. We have no competing interests.

Funding. This research was funded by NERC (NE/L501554/1 to J.E.O'R. and L.P.; NE/K500823/1 to M.N.P.; NE/L002434/1 to J.F.; NE/N003438/1 to P.C.J.D.), BBSRC (BB/N000919/1 to P.C.J.D.), the University of Bristol (STaR scholarship to A.R.T.), Royal Society Wolfson Research Merit Award (P.C.J.D.) and the John Templeton Foundation (43915 to D.P.).

Acknowledgements. We thank the other members of the Bristol Palaeobiology research group for discussion; Rob Asher (Cambridge) and Thomas Guillerme for comments on the draft manuscript. We also thank April Wright and an anonymous reviewer for their help improving the manuscript.

References

1. Harvey P., May, R, Nee, S. 1994 Phylogenies Without Fossils. *Evolution*. **48**, 523–529.
2. Rabosky DL. 2010 Extinction Rates Should Not Be Estimated From Molecular Phylogenies. *Evolution***. 64**, 1816–1824. (doi:10.1111/j.1558-5646.2009.00926.x)
3. Losos, JB. *et al.* 2013 Evolutionary Biology for the 21st Century. *PLoS Biol*. **11**, e1001466. (doi:10.1371/journal.pbio.1001466)
4. dos Reis M., Donoghue PCJ, Yang Z. 2016 Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet*. **17**, 1–10. (doi:10.1038/nrg.2015.8)

5. Goloboff PA, Carpenter JM, Arias JS, Miranda-Esquivel DR. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)

6. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)

7. Wright AM, Hillis DM. 2014 Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *PLoS One* **9**. (doi:10.1371/journal.pone.0109210)

8. O'Reilly JE, Puttick MN, Parry L., Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PC.J. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett*. **12**, 20160081-. (doi:10.1098/rsbl.2016.0081)

9. Congreve CR, Lamsdell JC. 2016 Implied weighting and its utility in palaeontological datasets: A study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–462. (doi:10.1111/pala.12236)

10. Holton TA. Wilkinson M, Pisani D. 2014 The shape of modern tree reconstruction methods. *Syst. Biol.* **63**, 436–441. (doi:10.1093/sysbio/syt103)

11. Guillerme T, Cooper N. 2016 Effects of missing data on topological inference using a Total Evidence approach. *Mol. Phylogenet. Evol.* **94**, 146–158. (doi:10.1016/j.ympev.2015.08.023)

12. Sanderson, M. J. & Donoghue, M. 1996 The relationship between homoplasy and the confidence in a phylogenetic tree. Academic Press, San Diego, California.

13. Goloboff PA, Farris S, Nixon K. 2008 TNT, a free programm for phylogenetic analysis. *Cladistics* **24**, 774–786. (doi:10.1111/j.1096-0031.2008.00217.x)

14. Stamatakis, A. 2014 RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)

15. Ronquist, F. *et al.* 2012 Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)

16. Wright AM, Lyons KM, Brandley MC, Hillis DM. 2015. Which came first: The lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. *J. Exp. Zool. B. Mol. Dev. Evol.* **324**, 504-416.

17. Nguyen L-T, Schmidt HA, von Haeseler A, BQ Minh 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268-274.

18. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)

19. Hilton J, Bateman RM. 2006 Pteridosperms are the backbone of seed-plant phylogeny. J. Torrey Bot. Soc. 133, 119–168. (doi:10.3159/1095-5674)

20. Luo ZX, Gatesy SM, Jenkins FA, Amaral WW, Shubin NH. 2015 Mandibular and dental characteristics of Late Triassic mammaliaform Haramiyavia and their ramifications for basal mammal evolution. *Proc. Natl. Acad. Sci.* **112**, E7101–E7109. (doi:10.1073/pnas.1519387112)

21. Nesbitt S.J, Barrett PM, Werning S, Sidor CA, Charig AJ. 2013 The oldest dinosaur? A Middle Triassic dinosauriform from Tanzania. *Biol. Lett.* **9,** 20120949. (doi:10.1098/rsbl.2012.0949)

22. Sutton MD, Briggs DEG, Siveter D.J, Siveter DJ, Sigwart, JD. 2012 A Silurian armoured aplacophoran and implications for molluscan phylogeny. *Nature* **490**, 94–97. (doi:10.1038/nature11328)

23. Mooers AO, Heard SB. 1997 Inferring Evolutionary Process from Phylogenetics Tree Shape. *Q. Rev. Biol.* **72**, 31–54.

24. Shao KT, Sokal RR. 1990 Tree balance. *Syst. Zool.* **39**, 266–276. (doi:10.1007/s13398-014-0173-7.2)

25. Harcourt-Brown K, Pearson P, Wilkinson M. 2001 The imbalance of paleontological trees. *Paleobiology* **27**, 188–204.

26. Panchen A. 1982 The use of parsimony in testing phylogenetic hypotheses. *Zool. J. Linn. Soc.* **74**, 305–328.

27. Scotland RW., Olmstead, R. G. & Bennett, J. R. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol.* **52**, 539–548. (doi:10.1080/10635150390223613)

28. Tarver, J. E. et al. 2016 The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biol. Evol.* **8**, 330-334. (doi:10.1093/gbe/evv261)

29. Healy, K., Guillerme, T., Finlay, S., Kane, A., Kelly, S. B., McClean, D., Kelly, D. J., Donohue, I., Jackson, A. L., and Cooper, N., 2014, Ecology and mode-of life explain lifespan variation in birds and mammals: Proc R. Soc. Biol Sci, **281**, 20140298.

465

	Asymmetrical generating phylogeny			
	Equal Weights Parsimony	**Implied Weights Parsimony**	**Maximum Likelihood**	**Bayesian**
100	**34.89 (22-56)**	**37.85 (22-56)**	**45.84 (20-58)**	**28.1 (18-39)**
350	**26.57 (11-51)**	**29.2 (12-51)**	**26.49 (6-58)**	**19.21 (7-35)**
1000	**17.82 (3-40)**	**19.16 (2-33)**	**11.94 (0-58)**	**9.34 (0-31)**
	Symmetrical generating phylogeny			
	Equal Weights Parsimony	**Implied Weights Parsimony**	**Maximum Likelihood**	**Bayesian**
100	**8.08 (0-33)**	**9.29 (0-29)**	**10.1 (0-58)**	**7.51 (0-29)**
350	**1.33 (0-28)**	**1.43 (0-28)**	**1.8 (0-52)**	**1.2 (0-28)**
1000	**0.32 (0-26)**	**0.31 (0-26)**	**0.51 (0-52)**	**0.31 (0-26)**

Table 1. Bayesian approaches produce the most accurate trees for all character sets. Mean and range (in brackets) of Robinson-Foulds distances are lower for topologies estimated using Bayesian methods for both the symmetrical and asymmetrical generating tree. Maximum Likelihood is the generally the most inaccurate method for the symmetrical generating tree, and implied weights parsimony performs worst for the asymmetrical generating tree.

	asymmetrical tree	symmetrical tree
MB 100	**< 0.001**	**0.09919**
Maximum Likelihood 100	**< 0.001**	**0.027295**
EW 100	**< 0.001**	**0.106712**
IW 100	**< 0.001**	**0.092736**
MB 350	**< 0.001**	**0.638242**
Maximum Likelihood 350	**< 0.001**	**0.057809**
EW 350	**< 0.001**	**0.19683**
IW 350	**< 0.001**	**0.148108**
MB 1000	**< 0.001**	**0.256976**
Maximum Likelihood 1000	**< 0.001**	**0.085987**
EW 1000	**< 0.001**	**0.179186**
IW 1000	**< 0.001**	**0.287058**

Table 2. P Values from Spearman's rank correlation between the percentage of nodes being accurately reconstructed and their distance from the root. Nodes closer to the tips are significantly more likely to be accurately reconstructed in asymmetrical trees but this is not generally true for symmetrical phylogenies.

Figure 1. Contour plots of Robinson-Foulds distance against phylogenetic resolution, indicating the higher accuracy of Bayesian implementations against all other methods with data generated on the asymmetrical phylogeny. The spectrum of red to yellow, reflect lower to higher density of trees. As the number of characters increases all methods converge on the correct phylogeny, although Bayesian phylogenies are generally the least resolved. The other methods achieve higher resolution but at a cost of lower accuracy. Data generated on the symmetrical phylogeny shows similar patterns but with much less variance and higher accuracy for all iterations; this lack of variance means point estimates cannot be shown as density estimates.

Figure 2. Accuracy of nodes is higher for those closer to the tips in the asymmetrical trees. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anti-clockwise order for Bayesian, Maximum likelihood, EW-Parsimony, and IW-Parsimony at each node. Accuracy of reconstructions is significantly lower in the 100 character dataset (a), and increases in the 350 character (b), and 1000 character datasets (c).

Figure 3. Accuracy of nodes is high for all nodes in the symmetrical phylogeny. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anti-clockwise order for Bayesian, Maximum likelihood, EW-Parsimony, and IW-Parsimony at each node. Accuracy of reconstructions is high in each dataset size, but there is a non-significant increase in accuracy as dataset size increases (a-c).

Figure 4. Alternative phylogenetic reconstruction methods alter our understanding of evolution with empirical matrices. However, the relationship of fossil seed ferns from Hilton and Bateman [19] is changed according to implementation (a-d), although *Caytonia* remains as sister to angiosperms in all analyses. Alternative analyses change the taxonomic affinity of *Kulindroplax* from Sutton et al. [22] (e-h).

Figure 5. Alternative phylogenetic reconstruction methods produce generally congruent reconstructions of evolution with empirical matrices. For Luo et al. [20], the relationship between the haramiyids and multituberculates is largely unchanged across analyses (a-d). IW-Parsimony (g) and Bayesian analyses place *Nyasasaurus* as close to the earliest dinosaur (e) and IW-Parsimony places it close to the earliest diverging taxa (g), but EW-Parsimony and Maximum Likelihood place the taxa as a derived member of Dinosauria (f,h).

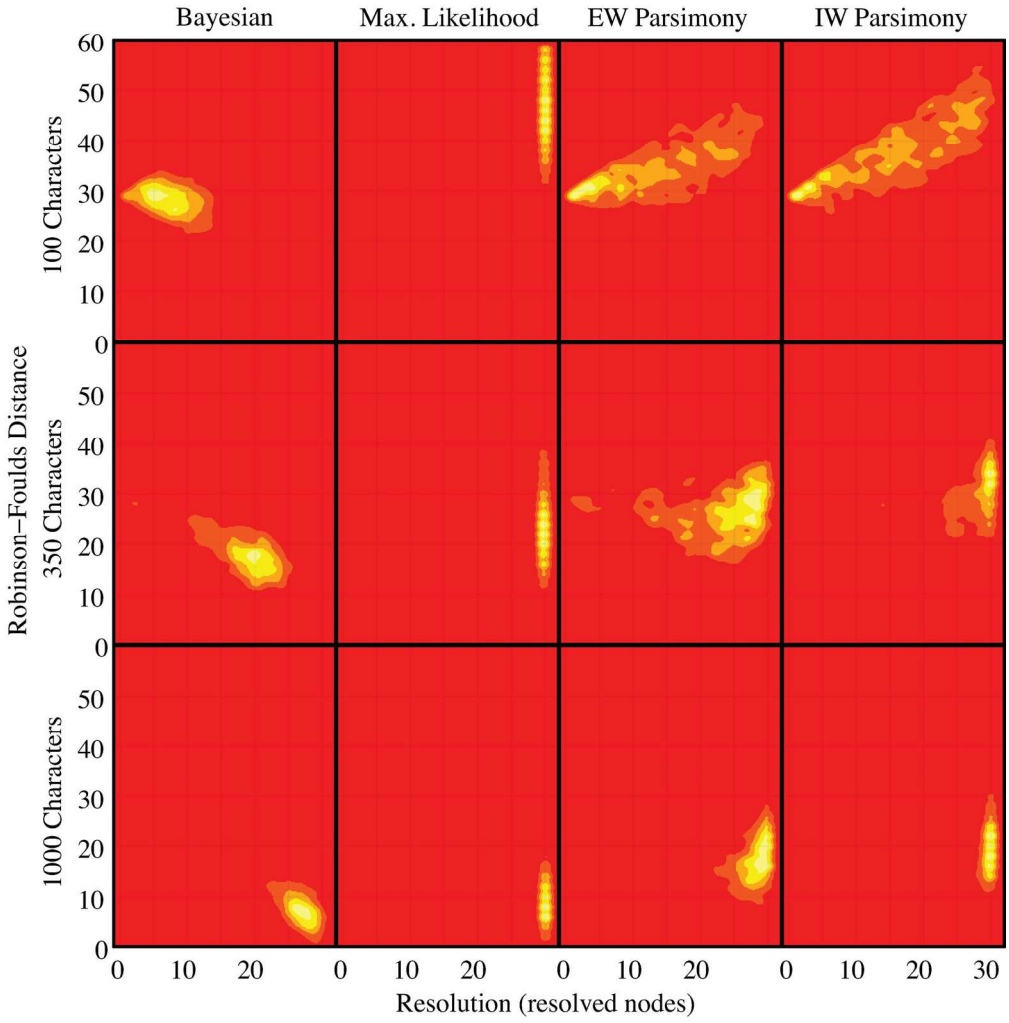

Figure 1

170x183mm (300 x 300 DPI)

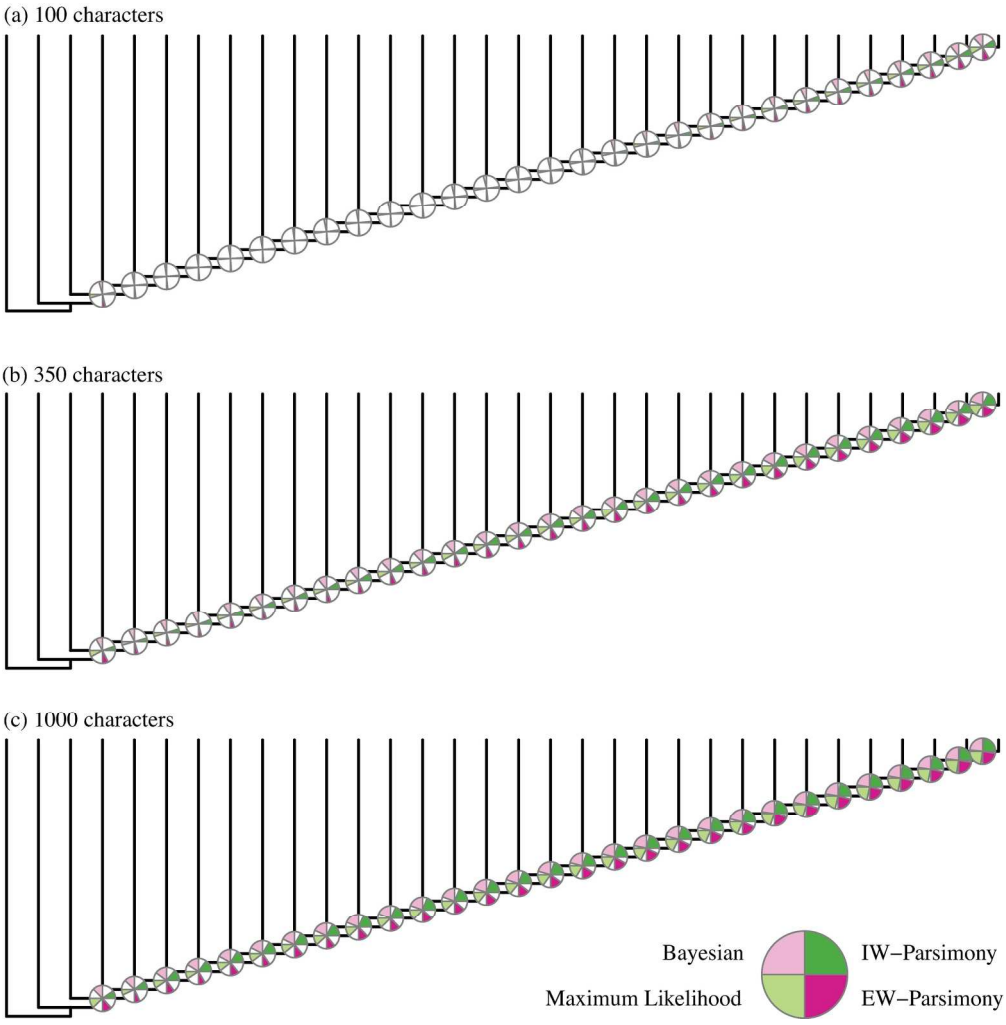

Figure 2

192x195mm (300 x 300 DPI)

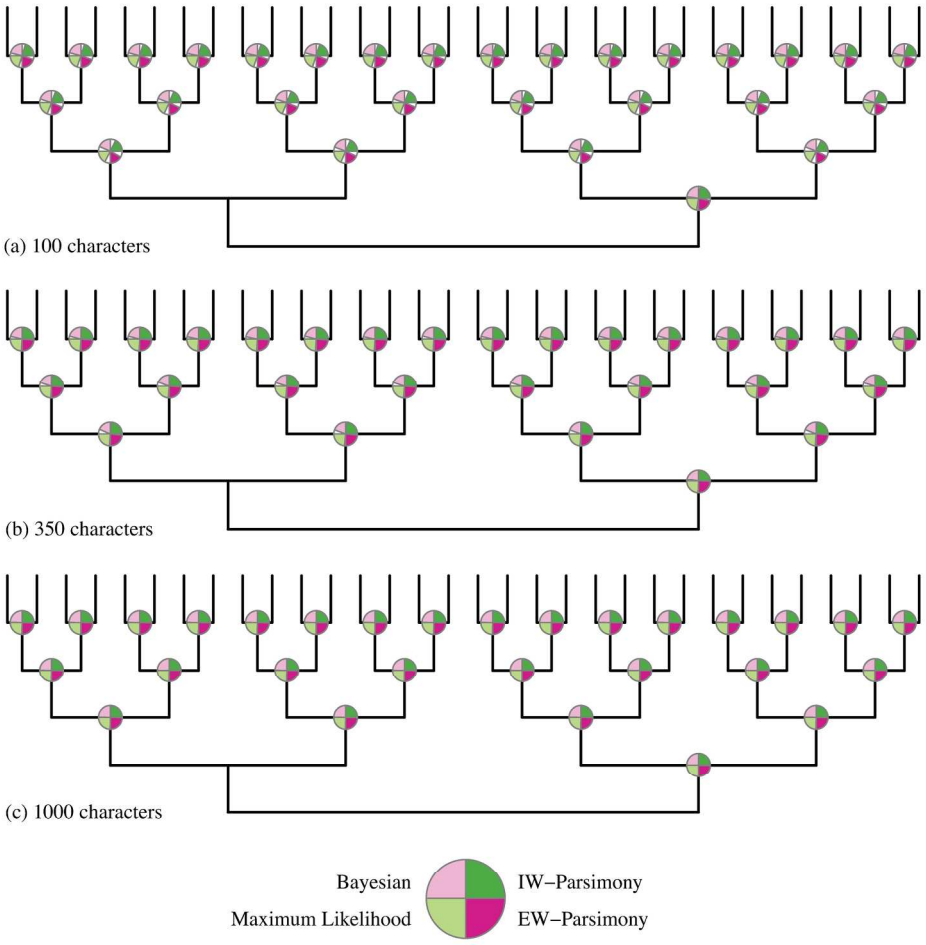

Figure 3

203x199mm (300 x 300 DPI)

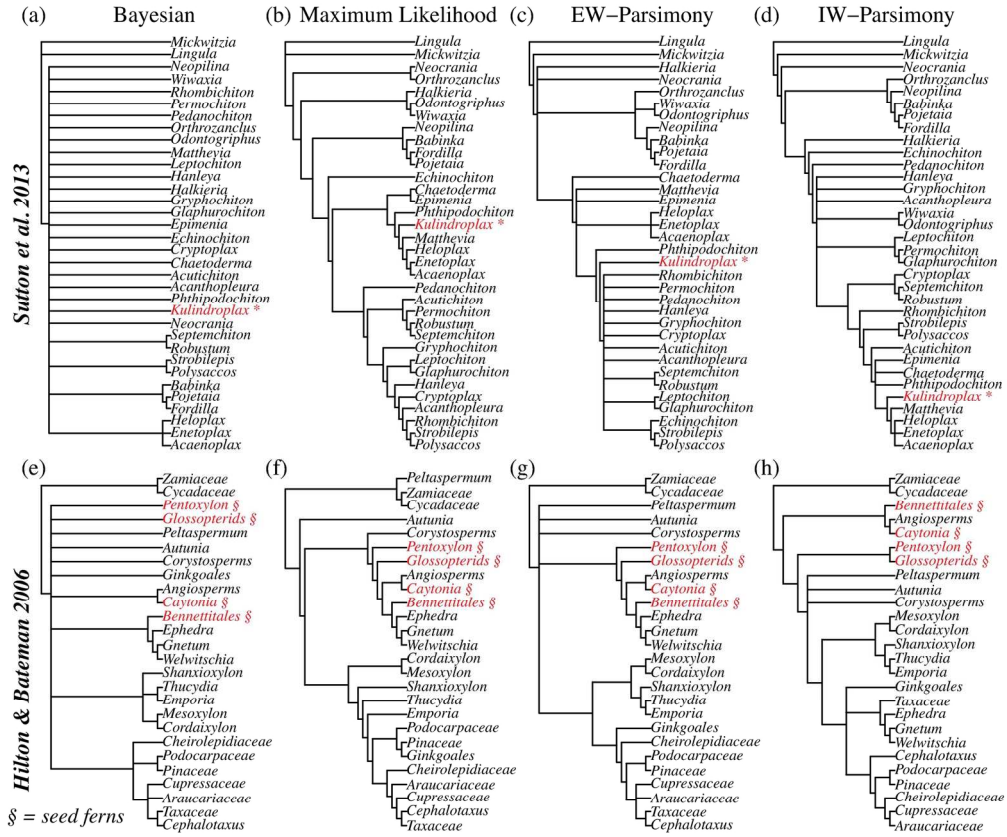

Figure 4

168x150mm (300 x 300 DPI)

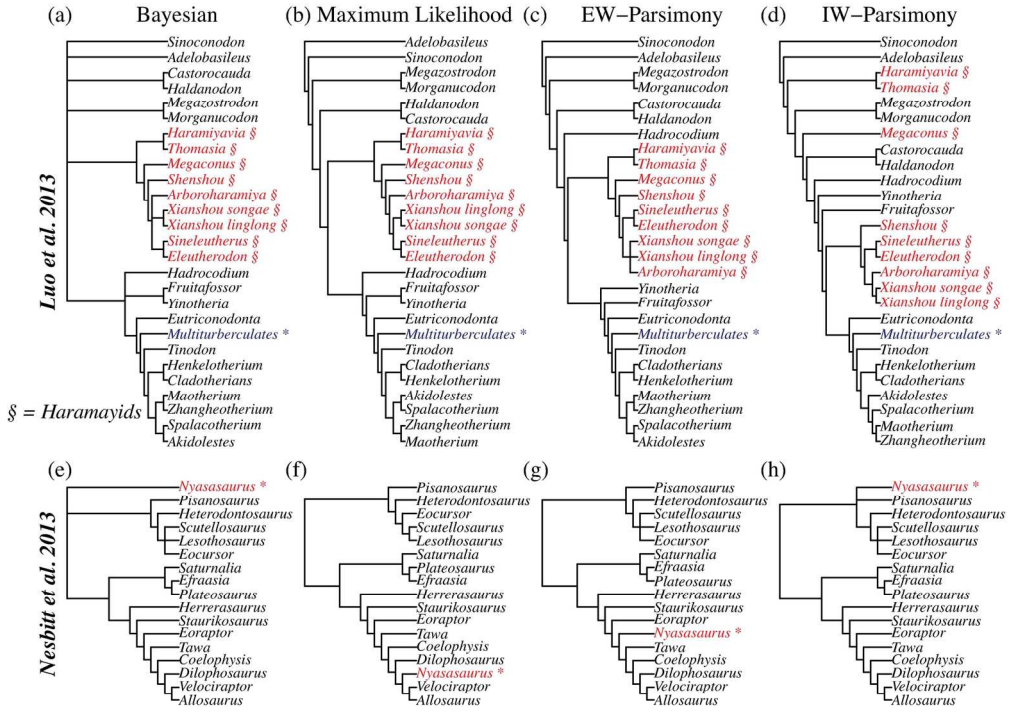

Figure 5

168x149mm (300 x 300 DPI)